

Fast Clusterwise m-Adic Regression: Application to Redshift calibration

Fionn Murtagh and Pedro Contreras

Department of Computer Science
Royal Holloway, University of London



Overview

- Introduction. The Baire metric concept
- Working with large data set, Berkeley Data Base (BDB)
- Applying Baire metric to astronomy data (SDSS)
- Data characterisation
- Clustering SDSS data based on a Baire distance
- Final remarks
- References

Introduction

Baire space consists of countable infinite sequence with a metric defined in terms of the longest common prefix [A. Levi. Basic set theory, Dover, 1979 (reprinted 2002)]

(The longer the common prefix, the closer a pair of sequence)

Consider two floating point numbers with the first p digits identical. Then what we call their Baire distance is 2^{-p} . This distance is an ultrametric. [see <http://www.cs.rhul.ac.uk/~fionn/papers>]

Introduction

It follows that a hierarchy can be used to represent the relationships associated with this distance.

We address the issue of whether such a hierarchy is advantageous, computationally, for clustering large data sets.

We seek to find inherent hierarchical structure in data, rather than fitting a hierarchy structure to data (as is traditionally used in multivariate data analysis).

We applied the Baire metric to Spectrometric and Photometric Red Shifts from the SDSS.

Baire, or longest common prefix

Definition: a Baire space consists of countably infinite sequence with a metric defined in terms of the longest common prefix.

Case of vectors x and y , with 1 attribute. Precision: digits 1, 2, ..., $|K|$

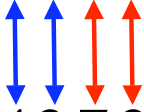
$$d_B(x_K, y_K) = \begin{cases} 1 & \text{if } x_1 \neq y_1 \\ \inf 2^{-n} & x_n = y_n \quad 1 \leq n \leq |K| \end{cases}$$

- each coordinate can be normalised (if needed), so is a floating point value.
- we define $d_B(x, y)$ based on sharing common prefix in all coordinates.

Baire, or longest common prefix

An example of Baire distance for two numbers (x and y) using a precision of 4

$x = 0.4256$
 $y = 0.4278$



Baire distance between x and y :

$$d_B(x_4, y_4) = 2^{-3} = |K| = 3$$

That is:

$$k=1 \rightarrow x_k = y_k \rightarrow 4$$

$$k=2 \rightarrow x_k = y_k \rightarrow 2$$

$$k=3 \rightarrow x_k \neq y_k \rightarrow 5 \neq 7$$

Working with large data sets

- Working with large data sets presents a challenge by itself.
- We seek computational advantage by using a Baire metric.
- We seek a cluster-wise regression, based on measurement precision, which can be useful in many cases (i.e. calibration).

Berkeley Data Base (BDB)

BDB is a general-purpose embedded database engine available as a set of computer libraries (APIs) to store and manage information.

BDB is:

- very fast
- highly configurable
- highly scalable (up to 256 terabytes, individual record up to 4 gigabytes of data)
- available in many programming languages
- designed for concurrent access

Berkeley Data Base (BDB)

- **Data persistence:** Once a data base has been created, data is persistent on that medium; this is opposite to storing information on RAM.
- **Querying:** BDB allows for a very simple matching and filtering, for example mathematical operators such as: +, -, <, =, >, etc. can be used without much worry about implementing data structures or program libraries to access this information efficiently.
- **Efficient access:** In our case a Java program was created to analyse and filter the SDSS data. BDB allows for access to the very same data base with programs written in C/C++ or several other languages.

These are of particular importance when working with large data sets where efficient filtering and retrieval is a must.

Berkeley Data Base (BDB)

BDB stores data in a B-tree key/value like structure not limited to scalar values (e.g. integer and strings); a value can be an arbitrary string of bytes, and a key can be any string of bytes that serves as unique identifier.

Working with Berkeley DB

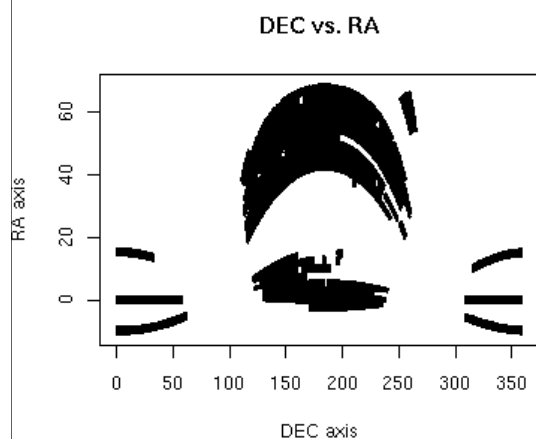
- **Environment:** a directory that encapsulates related databases and BDB infrastructure.
- **Database:** a collection of data items that share: index structure, a key and a set of secondary indices.
- **File:** contains one or more databases related to an environment.

BDB and Numerical Analysis

The first thing it is to export the data to BDB.

Having export our data to Berkeley data base we can start querying and analysing our data.

Data characterisation



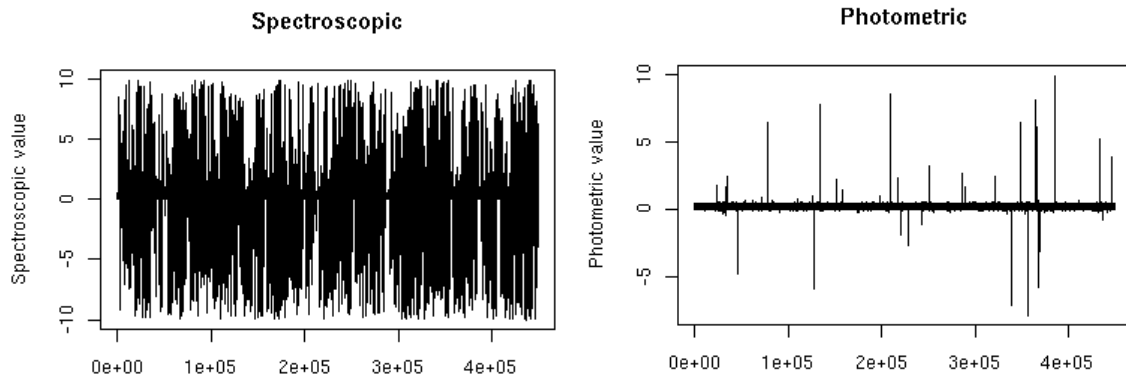
We have taken a subset of approximately 0.5 million data points from the SDSS release 5 [see D'Abrusco et al]:

Declination (DEC),
Right Ascension (RA),
Spectrometric,
Photometric

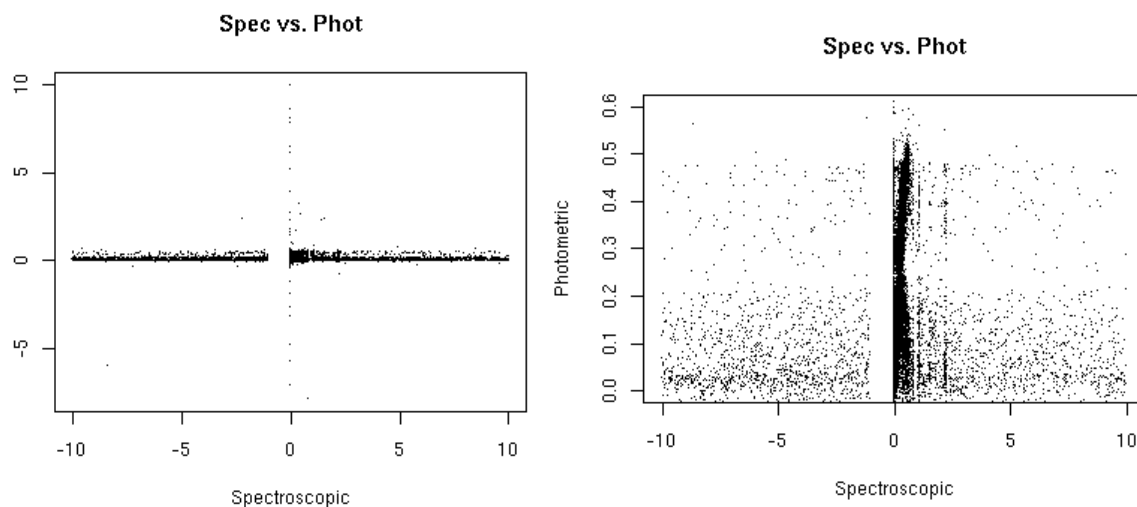
Dec vs RA are shown in the figure.

SDSS and Baire clustering

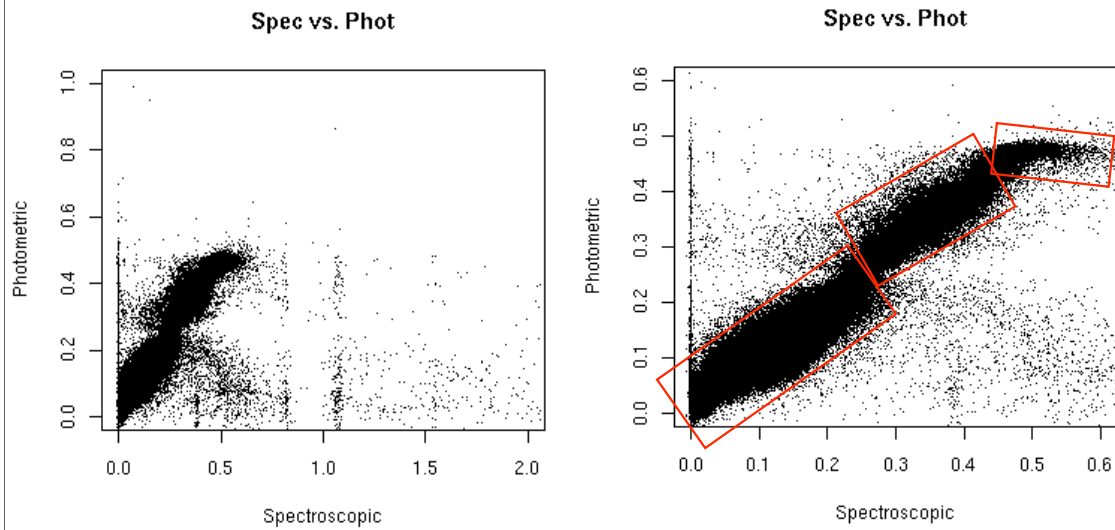
We seek to find inherent hierarchical structure in data, rather than fitting a structure to data (as is traditionally used in multivariate data analysis).



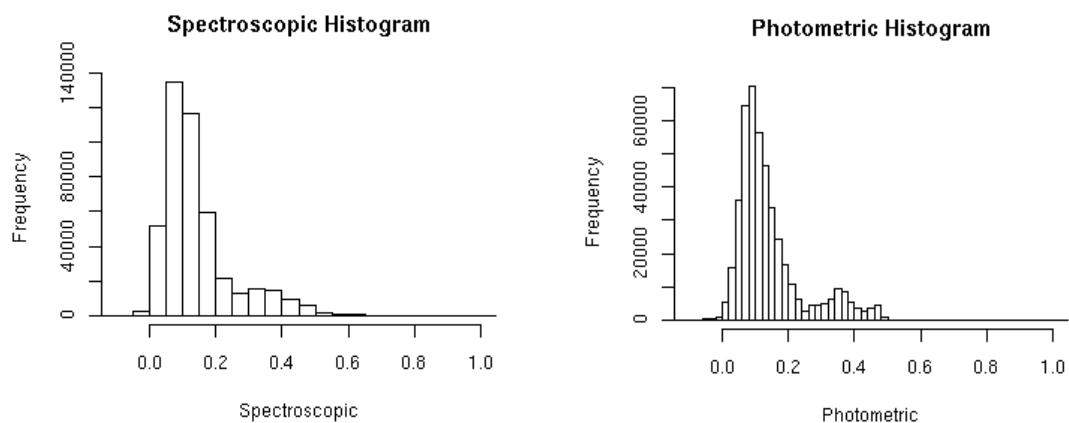
The data, Spec. vs. Phot.



Local linear fitting

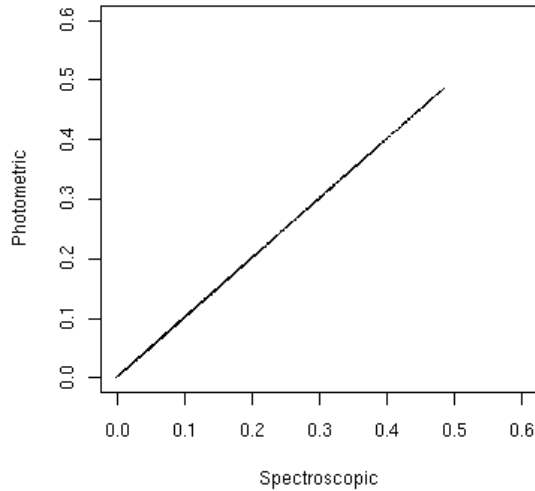


Density plots

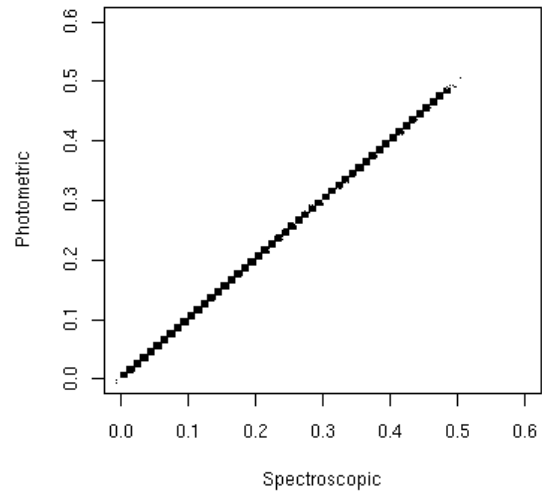


Second and third decimal digit

Spec vs. Phot, 3rd. decimal digit

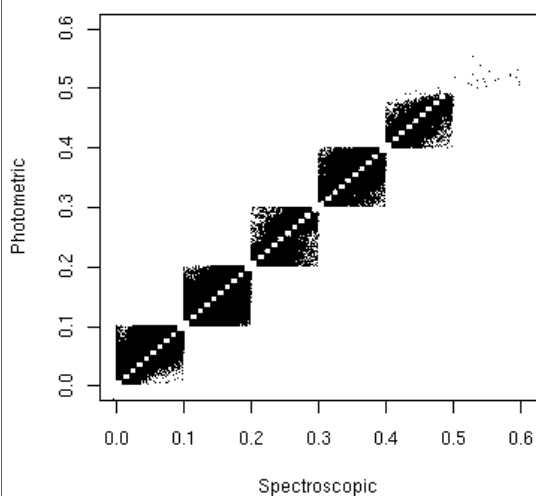


Spec vs. Phot, 2nd. decimal digit

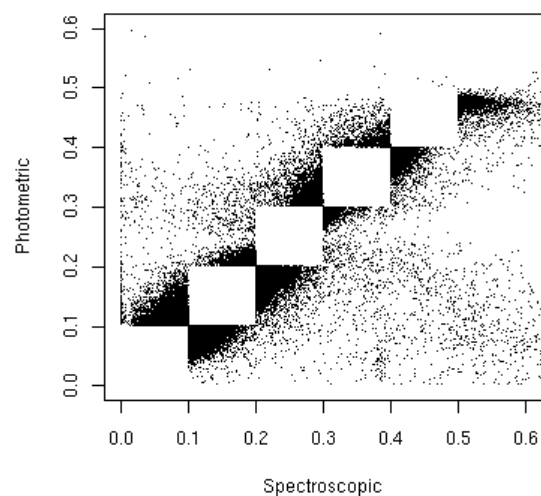


First digit and first decimal digit

Spec vs. Phot, 1st. decimal digit

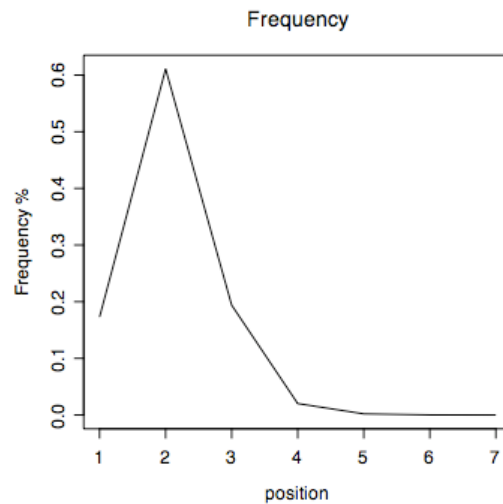


Spec vs. Phot, first digit



Prefix-wise clustering

prefix	No	%
1	76,567	17.28 %
2	270,497	61.06 %
3	85,911	19.39 %
4	8,981	2.02 %
5	911	0.02 %
6	90	0.0009 %
7	4	-
	442,961	100 %



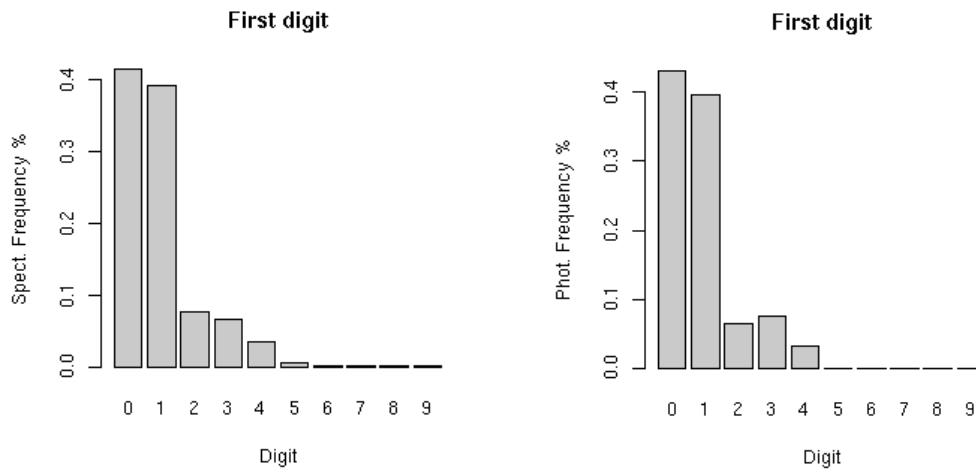
Benford's Law

Since we are working with digits precision we look into Benford's Law, to see if the data complies with it.

One may think that digits are distributed approx. equally, but it has been shown that digits follow the frequency distribution $\log(1 + 1/d)$.

Benford's Law has been used in a number of cases to detect data abnormalities and even fiscal fraud.

Benford's Law



Final remarks

- Based in prefix number precision we have shown that it is possible to find a mapping between the analysed data.
- Baire metric allows fast data clustering and processing
- BDB allows speed, scalability and flexibility when dealing with large data sets.

References

- F. Murtagh, G. Downs and P. Contreras, “Hierarchical Clustering of Massive, High Dimensional Data Sets by Exploiting Ultrametric Embedding”. SIAM Jnl. on Scientific Computing. Vol. 30, No. 2, pp. 707–730. February 2008.
- Oracle. Berkeley Data Base, 2008. <http://www.oracle.com/database/berkeley-db/index.html>
- Mining the SDSS archive. I. Photometric redshifts in the nearby universe. D’Abrusco et al. Draft version October 14, 2006.
- F. Murtagh, “Identifying the Ultrametricity of Time Series”, European Physical Journal B, 43, 573-579, 2005.
- F. Murtagh, “Hilbert Space Becomes Ultrametric in the High Dimensional Limit: Application to Very High Frequency Data Analysis”, arXiv:physics/0702064v1, submitted, 2007.
- F. Murtagh, “On Ultrametricity, Data Coding, and Computation”, Journal of Classification, 21, 167-184, 2004.
- F. Murtagh, “Thinking Ultrametrically”, in D. Banks, L. House, F.R. McMorris, P. Arabie and W. Gaul, Eds., Classification, Clustering, and Data Mining Applications, Springer, 3-14, 2004.
- F. Murtagh, “Quantifying Ultrametricity”, in J. Antoch, Ed., Compstat 2004: Proceedings in Computational Statistics, 1561-1568, Springer, 2004.
- Edouard Oblak, Bernard Debray, and Tomasz Kundera. BDB - A Database for All Types of Double Stars. In Mark G. Allen Francois Ochsenbein and Daniel Egret, editors, Astronomical Data Analysis Software and Systems XIII, volume 314, page 217. Astronomical Society of the Pacific, 2004.